Comment from Referee #1 (responses are in blue):

I find this article to be generally well-written, well-structured, and applies a transferrable methodology to classify stream conditions in ephemeral streams in a single study watershed. The authors support their claims and provide adequate figures to support their argument.

Thank you for your thoughtful review of this article.

I did find that some of the discussion sections strayed beyond the scope of the study described in the introduction section to discuss other features of the watershed and ephemeral streams more broadly. The paper would be strengthened by focusing on its central contribution.

These are valid points regarding the discussion sections, and we address them below.

I did find that a limitation of the study was that it focused on a subset of images from a single site. The methodology was demonstrated and its performance evaluated against predictions from the National Water Model, but statements about its transferability to other locations or are undercut by the limited nature of the data.

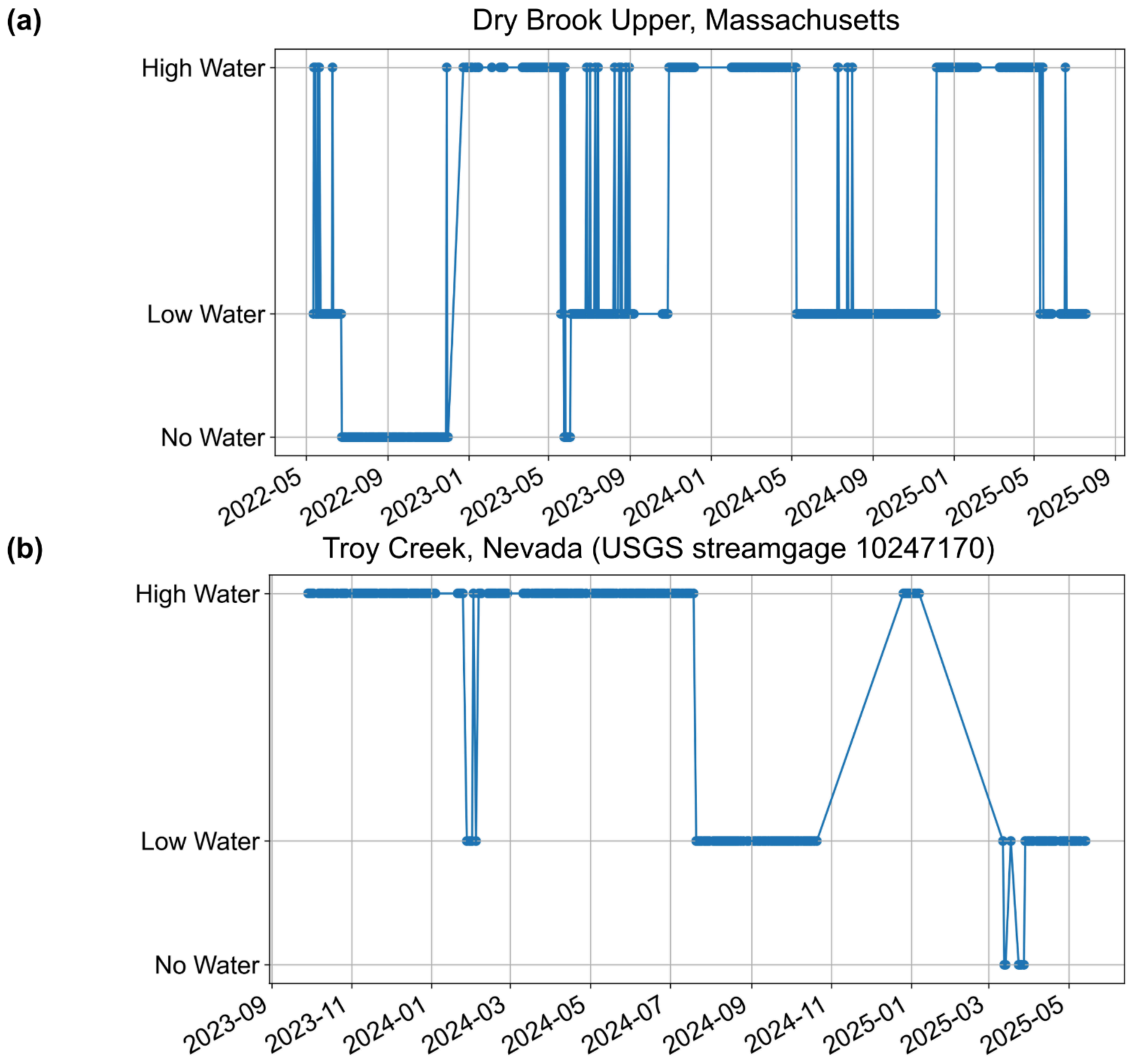
We appreciate this feedback and propose addressing it through the incorporation of an appendix that demonstrates transferability. The primary purpose of this work was to provide a proof of concept for this method using the Perry Creek site and its unique geographic context. While analysis of other sites in different locations and contexts would be beyond the scope of this paper, we can nevertheless provide a parsimonious set of examples to demonstrate the method's transferability.

We tested the model code (which is posted on the HydroShare repository associated with this paper) on two example sites from the USGS Flow Photo Explorer to produce time series of categorical flow states. These sites are Dry Brook Upper in Massachusetts and USGS streamgage 10247170 on Troy Creek in Nevada. Below, we have included a draft of the proposed new appendix text and figures, which would be referenced in a revised Discussion section, where extensibility and the USGS Flow Photo Explorer in particular are discussed.

Proposed appendix on transferability to other sites:

Although the main goal of this work was to demonstrate proof-of-concept at the PEC site, we also tested our model on two additional U.S. sites from the USGS Flow Photo Explorer: Dry Brook Upper in Massachusetts and USGS streamgage 10247170 on Troy Creek in Nevada (USGS, 2024). We selected these sites because they are both IRES with thousands of photos available. After labeling only 105 and 111 photos, respectively, the model achieved 84.4% accuracy at Dry Brook Upper and 76.5% at Troy Creek. The resulting time series of categorical flow states from model predictions (for all confidence levels) are shown in figure A12. This exercise was performed with fewer labeled photos compared to the PEC case, no photo cropping, and no changes to the model code (aside from updating the formatting of dates).

Based on this preliminary transferability analysis, we find that about 100 labeled images – with all categories represented in both training and testing sets – appear sufficient to transfer this method to other sites with consistent imagery. Notably, all photos used in this exercise were taken at noon, which likely enhanced model performance due to minimal variation in sun angle. While additional labeled photos would likely improve performance at any site, those with unbalanced categories or dramatic changes in illumination would benefit most.



Current deep learning models in the USGS Flow Photo Explorer (USGS, 2024) estimate relative flow states but cannot distinguish dry streambeds (Gupta et al., 2022; Goodling et al., 2025). Our method could complement the existing relative streamflow method, for example by being included in a conditional two-step approach: detect water presence first with our simple model; if present, estimate relative discharge using a CNN. This would preserve the simplicity and high accuracy of our model while enabling conditional estimation of streamflow when relevant. This approach would be well-suited to watersheds managed for both water supply and fishery health since both streamflow volume and stream connectivity would affect watershed management. With hundreds of thousands of photos available on the USGS Flow Photo Explorer (USGS, 2024) and the likelihood of increased IRES prevalence with climate change, this screening for IRES-specific states would be especially valuable. Instructions for applying our methodology to USGS Flow Photo Explorer images are available on HydroShare (see Data and Code Availability statement).

I have minor comments regarding clarity and a few considerations not in the original text but overall find the article a suitable contribution to HESS:

Thank you, I will address these comments below.

1. **Page 7, line 155:** What defines “environmental damage”? Tampering? Batteries dying?

   “Environmental damage” refers to various issues that can affect a field camera, such as the camera breaking due to water damage and the batteries dying due to the solar panel not receiving enough sunlight to charge fully. We propose editing the relevant text to: “... environmental damage, such as water damage or dirty solar panels not generating enough power.”

2. **Page 9, Lines 173-180**: The National Water Model (NWM) is trained/calibrated to gage flows, how close is the closest calibration site? In figure A1 looks like it is on the East Fork of Russian River, so not on the stream you are monitoring. Worth pointing out in this section.

   Yes, the nearest calibration site is the East Fork Russian River gage (EFR), which is not part of the Perry Creek watershed, but is part of the Lake Mendocino watershed, and is indicated in Figure A1. We propose adding the following sentence to Section 2.1: “Although the NWM was not calibrated using data from the Perry Creek watershed, it was calibrated using data from the USGS East Fork Russian River streamgage (EFR in Fig. A1; Cosgrove et al., 2024), also located within the Lake Mendocino watershed.”

3. **Figure 5:** You need axis labels indicating which axis is predicted and which is observed.

   Thank you for this suggestion. We will revise the figure to include the labels “Predicted Category” on the horizontal axis and “Observed Category” on the vertical axis.

4. **Page 9, line 197:** Indeed, cropping vegetation may be helpful here – if there is a mediterranean climate, vegetation dynamics and streamflow ephemerality are both highly seasonal the model could learn more from the banks (which could make up more of the image) than the channel where intended.

   This is an interesting insight, although its consideration – and the potential for bank vegetation to provide useful information for flow prediction (rather than reduce model performance, as in our case) – is beyond the scope of the present analysis. Nevertheless, we did not crop the images used in the new appendix application of photos from the USGS Flow Photo Explorer (see above), and we propose incorporating the following point into a revised Discussion section as follows: "Due to there being multiple field camera angles at PEC, we cropped the images to focus on the stream channel and staff plate. However, images do not necessarily need to be cropped, and bank vegetation could potentially help the model predict flow states. For sites with consistent camera types and viewing angles, a useful exercise could be to find the optimal image resolution and cropping extent for feature recognition. In such an exercise, the cost of increased computing power for higher resolution images should be balanced with model performance."

5. **Page 10, line 204:** You only labelled 12.8% of the total images you had available – this is acceptable but is a relatively small dataset for training or reporting performance (your testing set is 3.9% of your total image dataset) that will represent the population. This is a limitation of the study, since as you note the lighting can be very different at different times of day/year. Ideally you have a big enough testing set to represent performance at each class during different lighting (and vegetation and channel) conditions.

   Yes, we agree that the number of labeled photos is relatively small. Our intention, in part, was to demonstrate that the model can still perform well even with a limited number of classified photos; this is a situation that is common in data-limited settings. We believe this point is conveyed sufficiently throughout the paper and in our demonstration of strong model performance despite the limited number of labeled photos.

6. **Page 10, line 203-210:** Random sampling was used in the selection of images for training/testing, which is acceptable, but this means the performance is only representative of historical conditions coincident with the label dataset. The performance reported in this paper is not representative of model prediction on new unseen imagery. This point is worth noting to make sure the reader knows what the model performance represents.

   This is correct. Because the present study is limited to imagery from the study period, even "out-of-sample" testing data fall within the study period domain. Broadly, this means that model performance is only representative within the study period. However, to the extent that variation within the study period reflects variation outside of it, model performance during the study period may reasonably be considered indicative of performance under true out-of-period conditions. We thank the reviewer for raising this important point and propose incorporating the following language into Section 2.2.1 (and/or elsewhere, as appropriate): “Because this study is limited to imagery from the study period, our analysis and modeling strictly reflect that period. However, if the variation in imagery and corresponding flow during the study period captures the seasonal and inter-annual variability typical of other years, then the selected images may be considered broadly representative. In our case, the study period includes the full range from wet to dry years and thus arguably captures this variability.”

7. **Page 12, lines 247-250:** Why were these manual weights selected?

Thank you for noting the omission of our reasoning behind the selection of manual weights. We propose adding the following explanation to this section: "We assigned manual weights to emphasize water presence categories (‘high water’ and ‘low water’) over ‘no water,’ and gave the ‘obstructed’ category a weight higher than ‘no water’ -- reflecting its smaller sample size -- but lower than the water categories, given its lesser importance."

8. **Page 14, line 298:** Is there a reference for the 0.028 $m^3 s^{-1}$ threshold for NWM flow? How sensitive are your results to this selection? The selection of the threshold appears arbitrary at the moment.

   Thank you for noting this error. We ultimately did not use this threshold. As such, we propose replacing the line with the following description of what our analysis did: “For example, we calculated how often the observed stage at PEC was zero while the NWM predicted flow.”

9. **Figure 7:** Why are there negative stage values? And why are there purple high water dots in panel 7 when stage is reported negative? Is that supposed to be a diagnostic tool for quality assurance of the stage data, which leads to the record in panel b? The paragraph in the main text where Figure 7 is mentioned does not walk the reader through this. Also in Figure 7 are the stage observations without any dots times where there was no imagery or times where the imagery classification was deemed not high confidence? Consider adding shading to indicate “no imagery available” and another color of dot to indicate “no high confidence prediction” or something similar so the absence of data is clear.

   Thank you for your comments, which indicate that the placement and discussion of this image were not clear in the manuscript. To address this, we propose moving Figure 7 to the position of Figure 8, so that the relevant methods are discussed before the figure is presented. Otherwise, answers to most of the reviewer’s questions are already provided in Section 3.3 and Appendix 2. For remaining questions and omissions, we propose making clarifications in both the figure caption and the surrounding text. Specifically, we propose revising the figure 7 caption to read: “Original and quality-controlled, barometrically compensated stage from the Perry Creek (PEC) site from December 2022-February 2023. No imagery was available after 1 February 2023. Stage values (black lines) are colored (points) by high-confidence image classifications (only). a) Shows the time series before quality control, and b) shows the time series after quality control.” In addition, we propose emphasizing in Section 3.3 that observed stage data can be “prone to sensor error”. Finally, we propose adding text describing the negative stage values in more detail in Section 3.3: “Noisy data and stage measurements less than zero were an issue before installing the HOBO MX2001-04-SS-S pressure transducer and HOBO MicroRX data logger in October 2023; thus, the image classifications were useful in validating when the observed stage should be zero.”

10. **Page 28, line 446-448 and Page 29, line 464-465:** Is there a citation or the claim of not having enough imagery to train a CNN? The Gupta et al. and Noto et al. studies you cite have about as much imagery as you do. You report ~4,700 images, which is more than at 2 of Gupta et al. 's sites.

    We agree that our language misstates the point and is overly general. Gupta et al. found that using a reduced number of *labeled* image pairs (500–1,000) resulted in worse performance. In our study, we used 537 labeled images, even though the total number of available images -- both in our case and in Gupta et al.'s -- was higher. We intentionally limited the number of labeled images to evaluate model accuracy under more constrained conditions. Accordingly, we agree that the discussion of CNNs is not particularly helpful, as we did not explicitly evaluate CNN performance or its relationship to training size in our study. Therefore, we propose removing the references to CNNs and their sample size requirements.

11. **Section 4.4:** This section largely diverges from the central contribution of the study (a methodology to classify images) and into a lot of site-specific information that is largely conjecture about processes and reads as redundant to the prior section (4.3). This section could be eliminated.

    Thank you for your input. Upon review, we agree with your recommendation. We propose moving this section to an appendix and referencing it in the main text where appropriate.

12. **Page 33, line 604:** Is there a citation for the claim that "these efforts have struggled to translate to IRES"? Neither study cited included nonperennial streams.

    We agree, this should be reworded to: "these efforts did not focus on IRES".

13. **Section 4.6:** This section is only loosely connected with the central contribution of the paper (image classification model) and is material that could be included in the introductory material. This section could be eliminated.

    We agree that the material in this section is better suited for partial incorporation into the introduction, as well as inclusion in a new "Conclusion" section (in response to your comment below).

14. **Conclusion section is missing:** It is traditional to summarize the paper's contribution in a conclusion section, one is missing here.

    We propose including a new Conclusion section that summarizes the overall contribution, and incorporates salient points from Section 4.6 (in response to your comment above):

    "This work demonstrates that a simple machine learning algorithm can classify timelapse field camera images to identify no, low, or high water levels in IRES, providing a low-cost, transferable method for monitoring water occurrence in these sparsely observed systems. Given the prevalence of ungaged IRES, field cameras and image classification offer a practical approach to improving understanding of their role in climate-impacted freshwater systems. For example, the FIRO program at Lake Mendocino (Fig. A1) currently uses streamflow observations from EFR to inform reservoir inflow models. As climate change is expected to increase drying in IRES, unmonitored contributions from tributaries such as Perry Creek (Appendix 3) could affect reservoir storage. Thus, as the FIRO program expands, field cameras and image classification may offer a cost-effective approach to integrating information on the presence and magnitude of IRES contributions.

This approach can also support monitoring of critical habitats, including tributaries where salmon passage depends on streamflow connectivity threatened by drought and water diversions (see e.g., Scott River, 2025). Installing cameras at tributary confluences could inform targeted habitat restoration. More broadly, formally recognizing IRES as integral to river systems can incentivize monitoring and protect them from degradation due to climate change, mining, and urban development (Acuña et al., 2014). The 2023 exclusion of ephemeral streams from U.S. Clean Water Act protections highlights the vulnerability of IRES and the importance of cost-effective monitoring approaches like ours for understanding the impacts and effectiveness of water management efforts.

We conclude by offering practical recommendations for implementing our method. Cameras should be installed along IRES reaches that are important for monitoring water management objectives (e.g., fish passage, drought contingency planning). Installations should be in stable locations with clear views of the streambed and minimal vegetation interference. Consistent camera types and viewing angles should be used, as they improve the robustness of time series and the effectiveness of classification. Long-term maintenance budgets are also recommended to support sustained monitoring. Finally, this approach can be integrated with complementary methods (Gupta et al., 2022; Goodling et al., 2025) and deployed through accessible platforms such as the USGS Flow Photo Explorer (USGS, 2024) and the CrowdWater mobile application (SPOTTERON GmbH, 2025)."